EPIWORK FFCUL/LASIGE Progress Report

Mario Silva, Fabrício A. B. Silva, Francisco Couto, Luís Filipe Lopes, Patrícia Sousa, João Zamite, Hugo Ferreira

University of Lisbon, LASIGE - FFCUL

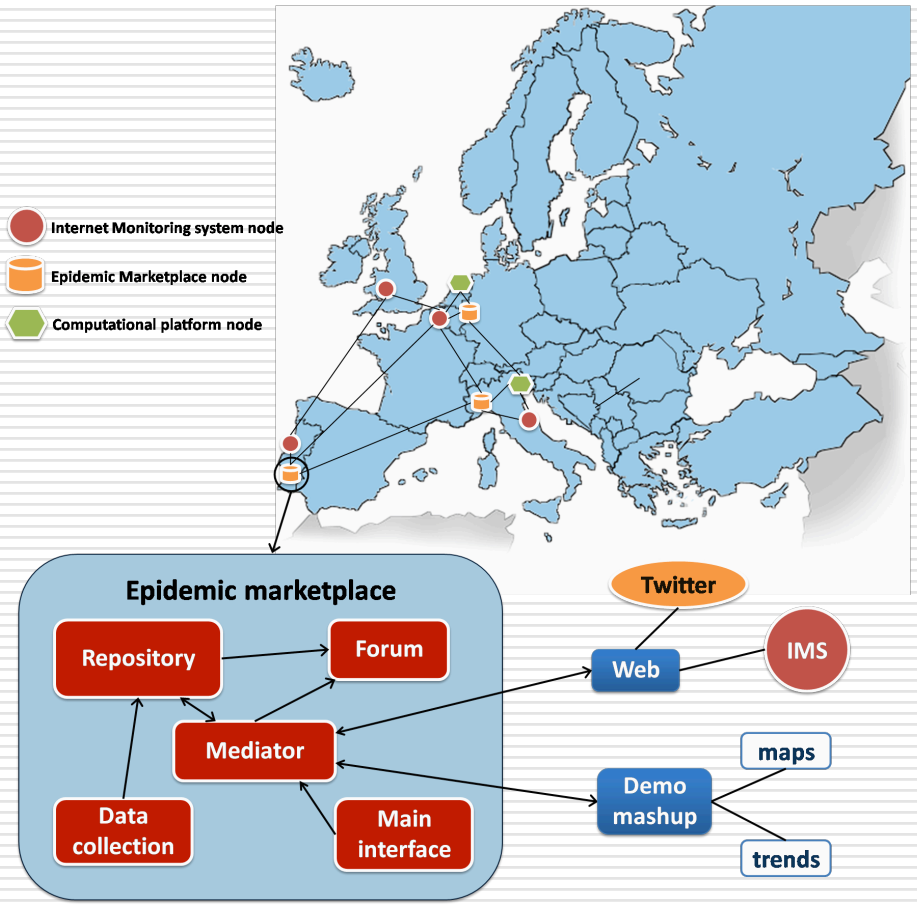
Advances and Contributions

- Work on WP3 – Epidemic Marketplace
- Deliverable D3.1 (meta-model) released
- Hardware and base software deployed
- Initial prototype of EM ready to receive your datasets

Epidemic Marketplace

- A **data integration platform** where epidemiological data can be stored, managed and made available to investigators, fostering collaboration.
- The EM has four main components (all currently under development):
  1. Data Collector
  2. Repository
  3. Middleware Services
  4. Forum

Epidemic Marketplace Architecture



Epiwork D3.1

Epiwork D3.1: Meta-model Initial Specification, Catalogue of Relevant Data, Platform Requirements

Luis F. Lopes, Fabricio Silva, Francisco Couto, Mario Silva

Common Reference Model

- Information to be described as **metadata**
  - Property-lists describing the epidemic datasets stored in the marketplace
  - Level of detail is key design point
- Must be/become **machine-readable**
  - Discoverable
  - Searchable
  - Accessible

Ontologies and Metadata

- Metadata to describe data
- Ontologies to describe relationships between data
  - Resolve ambiguities
  - Makes meta-data simpler

Metadata standards

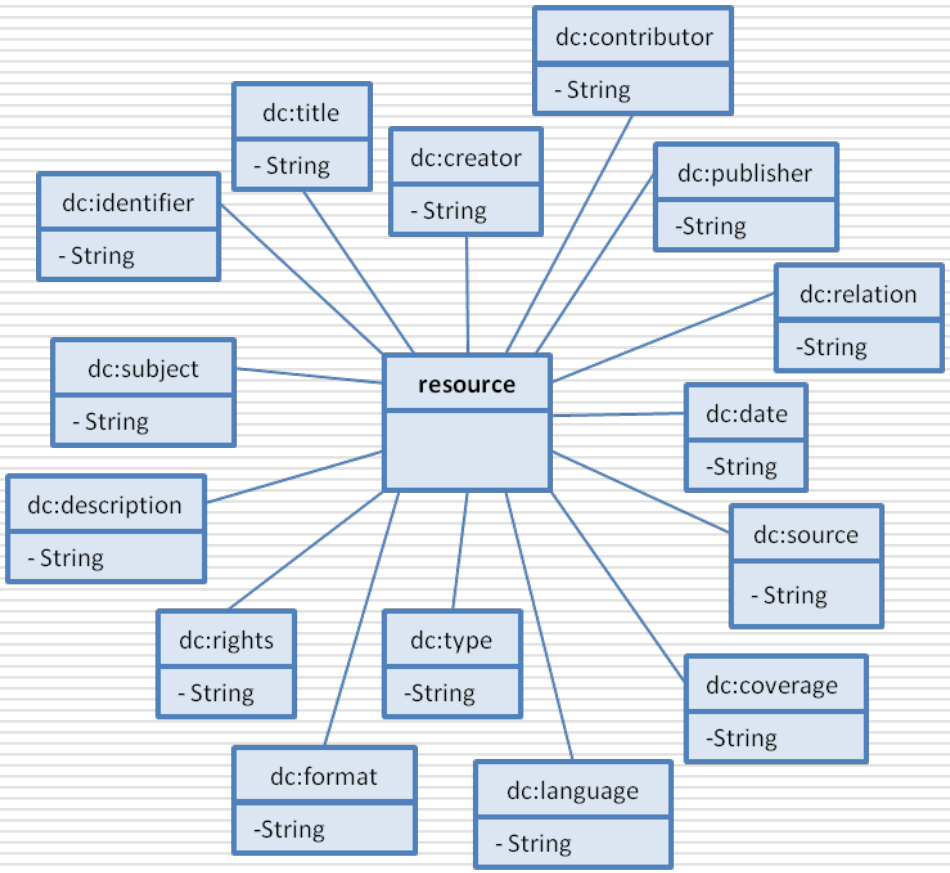
- ISO/IEC 11179 Metadata Registry (MDR)
- **Dublin Core (DC) metada for the Web, 15 properties**
  - ISO Standard Standard 15836-2003 of February 2003, ANSI/NISO Standard Z39.85-2007 of May 2007 and IETF RFC 5013 of August 2007.
  - **DCMI namespace**: Since 2008, DCMI includes formal domains and ranges in the definitions of its properties.

Ontology Standards?

- UMLS http://www.nlm.nih.gov/research/umls/
  - Too complex?
- OBO http://www.obofoundry.org/
- Inspire http://www.inspire-geoportal.eu/
  - Too complex?
  - Geonames, etc.

- Epidemiology is an open domain (will never be bounded)
- **First, we have to see the data.**

EM Catalogue – 1st. cut



Sample datasets (real)

- EM Twitter Datasets:
  - Twitter data harvested by an initial prototype of data collector module of the Epidemic Marketplace (Lopes et al. 2009)
- US Airports Dataset:
  - Data about the airport network of the United States (supplied by ISI)

Sample datasets (simulated)

- **Cohen et al. (2008)** – relationship between levels of household malaria risk and topography related humidity.
- **East et al. (2008)** - bird migration patterns; risk areas in Australia for avian influenza transmission from migrating birds.
- **Starr et al. (2009)** - Model for predicting the spread of Clostridium difficile in hospital context.

A twitter dataset

<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema- instance">

<**dc:creator**> LASIGE node of the Epidemic Marketplace</dc:creator>

<**dc:contributor**>Luís Filipe Lopes</dc:creator>

<dc:contributor>Joao M Zamite</dc:contributor> …

<**dc:coverage**>Spatial:Portugal</dc:coverage>

<**dc:coverage**>Temporal: 16-5-2009 to 3-6-2009</dc:coverage>

<**dc:language**>English</dc:language>

<dc:language>Portuguese</dc:language>

<**dc:source**>http://epiwork.di.fc.ul.pt/collector/</dc:source>

<**dc:identifier**>dataset-twitter-003</dc:identifier>

<**dc:date**>2009-05-29</dc:date>

<**dc:title**>Twitter dataset H1N1 + Portugal 4-6-2009</dc:title>

(to be continued)

</dc:dc>

A twitter dataset (2)

<**dc:subject**>twitter message dataset</dc:subject>

<**dc:format**>text/tab-separated-values</dc:format>

<**dc:type**>dataset</dc:type>

<**dc:description**> This dataset contains Twitter messages containing the words H1N1 and Portugal collected between 16-5-2009 and 3-6-2009, Information is a 7 columns relation, containing the following data:

- Column 1- keyword 1 (disease)- H1N1
- Column 2- Keyword 2 (location)- Portugal
- Column 3- Source (Twitter)
- Column 4- Author of the message (user id)
- Column 5- The message body (evidence)
- Column 6- score
- Column 7- date (day and hour)

</dc:description>

A twritter dataset (3)

<**dc:publisher** >Epiwork – http://www.epiwork.eu </ dc:publisher>

<**dc:relation**> Luis F. Lopes, João M. Zamite, Bruno C. Tavares, Francisco M. Couto, Fabrício Silva and Mário J. Silva. (2009). Automated Social Network Epidemic Data Collector. INForum informatics symposium.</dc:relation>

<**dc:rights**> Creative Commons Attribution-ShareAlike (CC BY-SA),http://creativecommons.org/licenses/by-sa/3.0/

</dc:rights>

US Airports dataset

<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<dc:contributor> Daniela Paolotti, ISI </dc:creator>

<dc:coverage>United States</dc:coverage> <dc:language> </dc:language>

<dc:source> </dc:source> <dc:identifier></dc:identifier>

<dc:format>text/plain</dc:format> <dc:date>2009-09-03</dc:date>

<dc:title> US Air Transportation Network </dc:title>

<dc:creator> ISI node of the Epidemic Marketplace</dc:creator>

<dc:subject> Undirected weighted network of the 500 US airports with the largest amount of traffic </dc:subject>

<dc:type>dataset</dc:type> <dc:description> </dc:description>

<dc:publisher >Epiwork – http://www.epiwork.eu </dc:publisher>

<dc:relation></dc:relation>

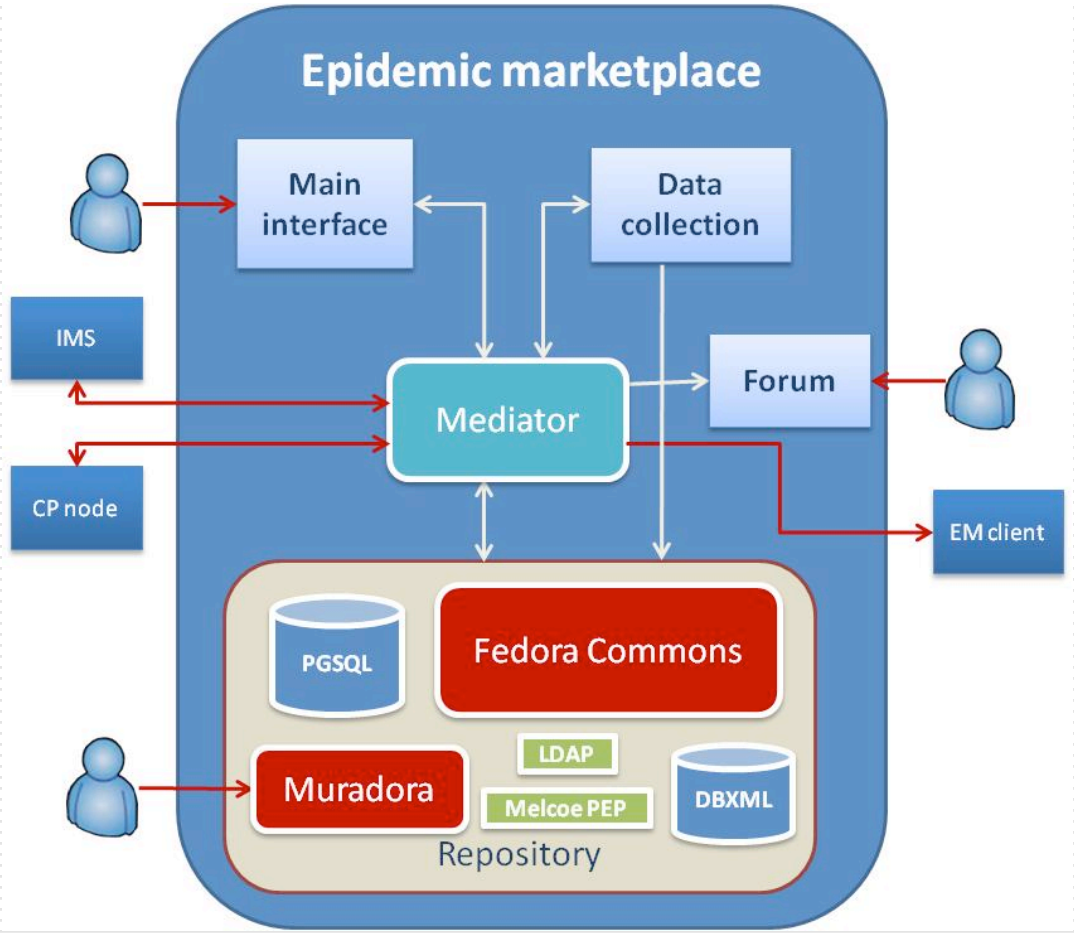
<dc:rights> Please, feel free to use the above network dataset, provided the appropriate credit is given to the authors </dc:rights>

</dc:dc>

APIs and Machine access

- PMH - Standard protocol/API for DC meta-data exchange
- ORE – SW style data constellations
- The Semantic Web stack

Architecture



Contact

Welcome fedoraAdmin | Your Account | Logout

EPIWORK

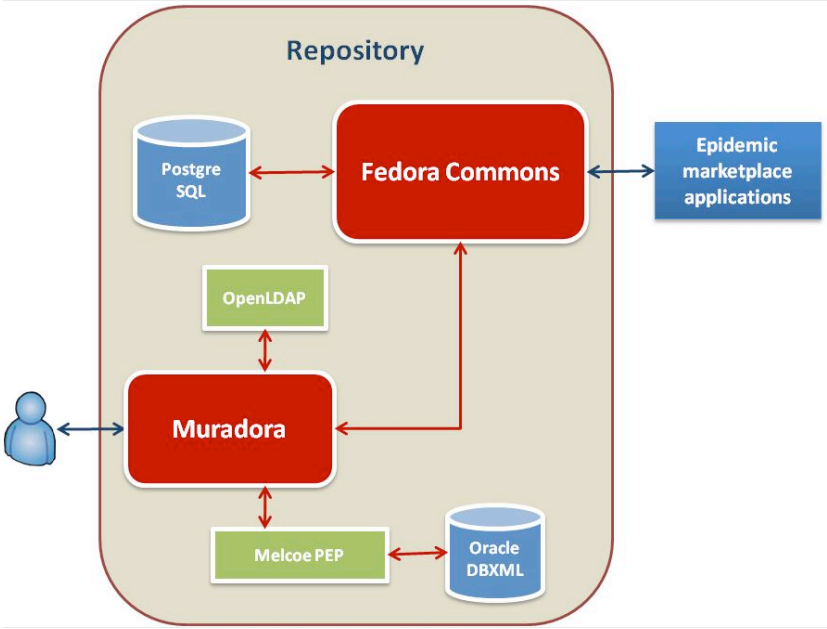
Home | Browse | Search | Submit | Publish | Portfolio | Admin Tools

Home

Resource Metadata

Meta-data	
Title	Twitter dataset H1N1 + Portugal 4-6-2009
Creator	LASIGE node of the Epidemic Marketplace
Subject	twitter message dataset
Description	This dataset contains Twitter messages containing the words H1N1 and Portugal collected between 16-5-2009 and 3-6-2009, Information is a 7 columns relation, containing the following data: Column 1- keyword 1 (disease)- H1N1 Column 2- Keyword 2 (location)- Portugal Column 3- Source (Twitter) Column 4- Author of the message (user id) Column 5- The message body (evidence) Column 6- score Column 7- date (day and hour)
Publisher	Epiwork – http://www.epiwork.eu
Format	text/tab-separated-values
Language	English, Portuguese
Contributor	Luis F Lopes, Joao M Zamite, Bruno C Tavares, Francisco M Couto, Fabricio Silva, Mario J Silva
Relation	Luis F. Lopes, João M. Zamite, Bruno C. Tavares, Francisco M. Couto, Fabrício Silva and Mário J. Silva. (2009). Automated Social Network Epidemic Data Collector. INForum informatics symposium.
Source	http://epiwork.di.fc.ul.pt/collector/
Coverage	Spatial: Portugal, Temporal: 2009-5-16 to 2009-6-3
Rights	Creative Commons Attribution-ShareAlike (CC BY-SA), http://creativecommons.org/licenses/by-sa/3.0/

Repository– Initial Deployment



- Web services interface to Fedora commons
- LDAP user registry
- Policy enforcement point (PEP) and XACML role-based access control
- Muradora front-end likely to go

What data?

Good

- Some dimensions are common
  - Geographic
  - Demographic

No Good

- Every epidemiologic study has its specific dimension, so do epidemic models

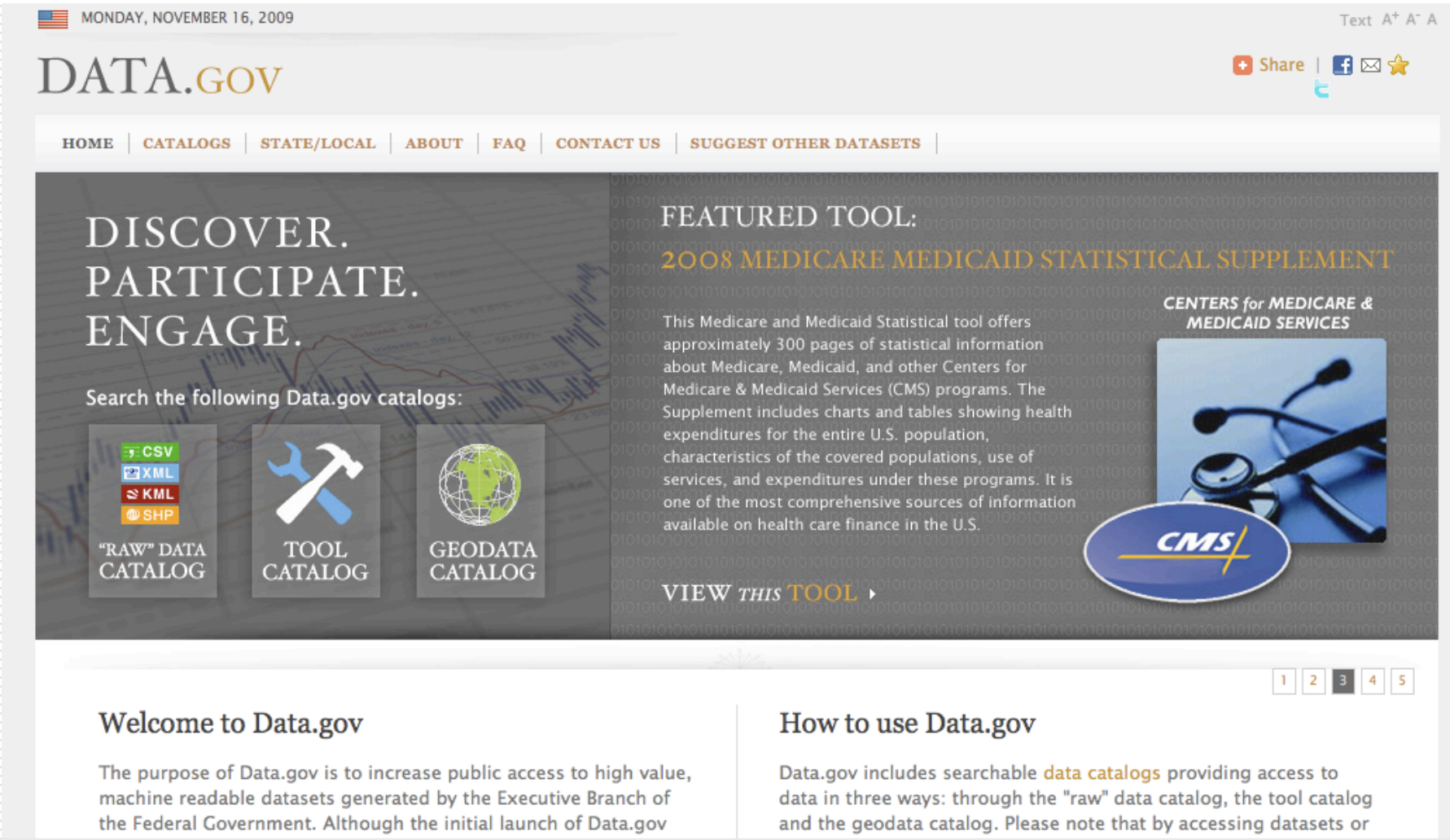
How do we get the data?

- Providers
  - Public administration authorities
  - Data collection organizations ($-$$$)
- Collectors
  - WP5
  - APIs, protocols?
- Modellers
  - Help on identifying data
  - Results!

How do we distribute the data?

- Not a problem of the 1st year
- Issues
  - Anonymization
  - Access control policies
  - Selection and dataset generation
  - Integration with computational platform

Data.gov


MONDAY, NOVEMBER 16, 2009
Text A+ A- A
DATA.GOV
Share
HOME | CATALOGS | STATE/LOCAL | ABOUT | FAQ | CONTACT US | SUGGEST OTHER DATASETS
DISCOVER.
PARTICIPATE.
ENGAGE.
Search the following Data.gov catalogs:
CSV
XML
KML
SHP
"RAW" DATA CATALOG
TOOL CATALOG
GEODATA CATALOG
FEATURED TOOL:
2008 MEDICARE MEDICAID STATISTICAL SUPPLEMENT
This Medicare and Medicaid Statistical tool offers approximately 300 pages of statistical information about Medicare, Medicaid, and other Centers for Medicare & Medicaid Services (CMS) programs. The Supplement includes charts and tables showing health expenditures for the entire U.S. population, characteristics of the covered populations, use of services, and expenditures under these programs. It is one of the most comprehensive sources of information available on health care finance in the U.S.
CENTERS for MEDICARE & MEDICAID SERVICES
CMS
VIEW THIS TOOL
1 2 3 4 5
Welcome to Data.gov
The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government. Although the initial launch of Data.gov
How to use Data.gov
Data.gov includes searchable data catalogs providing access to data in three ways: through the "raw" data catalog, the tool catalog and the geodata catalog. Please note that by accessing datasets or

Thank You!

- See
  - http://epiwork.di.fc.ul.pt/
- Attend
  - Today's demo on how to upload datasets
  - WP3+WP4 meeting tomorrow

WP3 SWOT Analysis

Strengths

- Epiwork-driven EM
- Standards-based
- Open Source modules
- Supported (until 2012)

Weaknesses

- **Unpopulated EM**
- What are the incentives?
- Interfaces to WP4 and WP5?
- Compelling scenario?

WP3 SWOT Analysis

Opportunities

- Epiwork testbed
- Creation of a baseline for epidemic modelling
- Showcase for partners' outputs

Threats

- Consortium enters "everyone for himself" mode.
- "Somebody will take care of that" attitude
- Someone turning EM into a very expensive complex and useless cache (not likekly)

Todo list and planning

BD - Before Deliverable

1. Populate Repository
2. Specify Middleware APIs
3. Release Collector
4. Replace Muradora?
5. Integrate Forum

AD - After Deliverable

1. Populate Repository
2. Ethics, Privacy and Anonimization
3. Distributed Authentication
4. Access control
5. Dataset selection generation
6. Replicate EM node?

Scheduled Deliverables

		Year 1			Year 2			Year 3			Year 4		
		M4	M8	M12	M16	M20	M24	M28	M32	M36	M40	M44	M48
WP3	**Information platform**												
Task 1			D3.1										
Task 2			D3.1										
Task 3				D3.2		D3.3				D3.4			D3.6
Task 4										D3.4 D3.5			D3.6

Scheduled Deliverables WP3

3.1	Report - Meta-model initial specification , catalogue of relevant data, platform requirements.	3	FFCUL	13 (9)	O+ R	CO	Month 8
3.2	Prototype of the Epidemic Marketplace Platform with an initial set of epidemiological databases integrated available to project participants.	3	FFCUL	39 (25)	P	CO	Month 12
3.3	Public release of the Epidemic Marketplace Platform.	3	FFCUL	26 (17)	D	PU	Month 20
3.4	Report for the Epidemic Marketplace Platform after release, describing changes and new implemented features, new data sources integrated with the platform, results of user surveys and usage statistics.	3	FFCUL	13 (10)	R	CO	Month 36
3.5	Report – Epidemic data ontology.	3	FFCUL	13 (10)	R	PU	Month 36
3.6	Report – Final specification of the Epidemic Marketplace Platform and evaluation results.	3	FFCUL	24 (18)	R	PU	Month 48